DEPARTAMENTO DE
INGENIERÍA
INFORMÁTICA
UNIVERSIDAD DE SANTIAGO DE CHILE

EJERCICIO PRÁCTICO 14: REGRESIÓN LOGÍSTICA

CONTEXTO

Conocemos varias herramientas que facilitan la búsqueda y construcción de modelos de regresión lineal, además del proceso iterativo para conseguir un modelo confiable.

El objetivo de este ejercicio es utilizar herramientas y procedimientos análogos para crear y evaluar modelos de regresión logística (RLog) para predecir una variable dicotómica.

OBJETIVOS DE APRENDIZAJE

1. Preparar un conjunto de datos para la construcción de modelos RLog.
2. Iterar en el proceso de selección de variables, creación y evaluación de modelos RLog, hasta conseguir uno que sea confiable y satisfactorio.

ÉXITO DE LA ACTIVIDAD

El equipo es capaz de encontrar modelos RLog confiables y de buen desempeño al predecir una variable dependiente.

ACTIVIDADES

Para esta actividad usaremos los datos de medidas anatómicas recolectados por Heinz et al. (2003) que ya conocimos en el ejercicio práctico anterior. Como este ejercicio requiere de una variable dicotómica, vamos a realizar lo siguiente:

- Crear la variable IMC (índice de masa corporal) como el peso de una persona (en kilogramos) dividida por el cuadrado de su estatura (en metros).
- Si bien esta variable se usa para clasificar a las personas en varias clases de estado nutricional (bajo peso, normal, sobrepeso, obesidad, obesidad mórbida), para efectos de este ejercicio, usaremos dos clases: sobrepeso (IMC ≥ 25,0) y no sobrepeso (IMC < 25,0).
- Crear la variable dicotómica EN (estado nutricional) de acuerdo al valor de IMC de cada persona.

1. El equipo copia el enunciado del problema asignado como comentarios de un script R.
2. El equipo lee el enunciado, descarga el archivo de datos (EP13 Datos.csv) desde UVirtual y selecciona las columnas para trabajar de acuerdo con las instrucciones.
3. El equipo construye los modelos solicitados usando la muestra correspondiente.
4. El equipo sube el script con las actividades anteriores comentando en detalle los pasos seguidos.

Fuera del horario de clases, cada equipo debe subir el script realizado UVirtual con el nombre "EP14-respuesta-grupo-i", donde i es el número de grupo asignado. Las respuestas deben subirse antes de las 23:30 del miércoles 8 de junio.

PREGUNTA (todos los grupos)

Ahora podemos construir un modelo de regresión logística para predecir la variable EN, de acuerdo con las siguientes instrucciones:

1. Definir la semilla a utilizar, que corresponde a los últimos cuatro dígitos del RUN (sin considerar el dígito verificador) del integrante de mayor edad del equipo.
2. Seleccionar una muestra de 120 mujeres (si la semilla es un número par) o 120 hombres (si la semilla es impar), asegurando que la mitad tenga estado nutricional "`sobrepeso`" y la otra mitad "`no sobrepeso`". Dividir esta muestra en dos conjuntos: los datos de 80 personas (40 con EN "`sobrepeso`") para utilizar en la construcción de los modelos y 40 personas (20 con EN "`sobrepeso`") para poder evaluarlos.
3. Recordar las ocho posibles variables predictoras seleccionadas de forma aleatoria en el ejercicio anterior.
4. Seleccionar, de las otras variables, una que el equipo considere que podría ser útil para predecir la clase EN, justificando bien esta selección.
5. Usando el entorno R y paquetes estándares[1], construir un modelo de regresión logística con el predictor seleccionado en el paso anterior y utilizando de la muestra obtenida.
6. Usando herramientas estándares[1] para la exploración de modelos del entorno R, buscar entre dos y cinco predictores de entre las variables seleccionadas al azar, recordadas en el punto 3, para agregar al modelo obtenido en el paso 5.
7. Evaluar la confiabilidad de los modelos (i.e. que tengan un buen nivel de ajuste y son generalizables) y "arreglarlos" en caso de que tengan algún problema.
8. Usando código estándar[1], evaluar el poder predictivo de los modelos con los datos de las 40 personas que no se incluyeron en su construcción en términos de sensibilidad y especificidad.

CRITERIOS DE EVALUACIÓN

Categoría	Nivel de logro	Puntos
Datos	Agregan la variable 'estado nutricional' (EN), estimada correctamente a partir de la variable 'índice de masa corporal' (IMC), al conjunto de datos y seleccionan muestras de entrenamiento y prueba siguiendo las instrucciones dadas y asegurando que esta variable se encuentra balanceada en ellas	3
Modelo de regresión logística simple (RlogS)	Seleccionan, justificando su utilidad de forma convincente, una variable de entre las no reservadas para explorar que utilizan para construir correctamente un modelo de RLogS para predicir la variable EN, evitando las variables obviamente correlacionadas	3
Selección de variables	Seleccionan un conjunto de variables relevantes para predecir la variable EN, utilizando correctamente gráficos y/o utilidades en paquetes de R para explorarlas, pero sin hacer uso de herramientas del paquete `caret`, desde el conjunto de ocho variables elegidas aleatoriamente en el ejercicio pasado y respetando las otras restricciones indicadas en el enunciado	4
Modelo de regresión logística múltiple (RlogM)	Construyen correctamente un modelo de RLogM para predicir la variable EN agregando las variables seleccionadas anteriormente al modelo RlogS que se tiene	2

[1] Entenderemos esto como paquetes tradicionales, sin incluir el paquete `caret`.

Confiabilidad de los modelos	Escriben comentarios y código en R correcto que verifica las condiciones que garantizan que tanto el modelo de RLogS como el modelo de RLogM obtenidos tienen un buen nivel de ajuste y son generalizables, interpretando explícita y correctamente los resultados obtenidos en cada paso y tomando acciones correctivas apropiadas de ser necesarias o comentando los riesgos asociados	4
Calidad predictiva de los modelos	Escriben código R correcto que evalúa la calidad predictiva, tanto del modelo de RLogS como del modelo de RLogM obtenidos, en datos no utilizados para su construcción, y comparando correctamente los desempeños observados	3
Conclusión	Entregan conclusiones correctas y completas, basadas en las evaluaciones realizadas y el proceso de búsqueda de predictores seguido, respecto del modelo de RlogM obtenido	2
Código fuente	El script está completo, ordenado y bien indentado, logrando un programa que es fácil de seguir, sin sentencias espurias y que no requiere cambios para que funcione	3
Ortografía y redacción	El script está comentado paso a paso, con claridad (basta una lectura para entender) y con buena redacción y ortografía (<= 5 errores), usando vocabulario propio de la disciplina y el contexto del problema	3
	TOTAL	**27**
	NOTA	**7,0**